CHAPTER 02

Collection and Organization of Data

Chapter Contents

You should read this chapter if you need to learn about:

Data

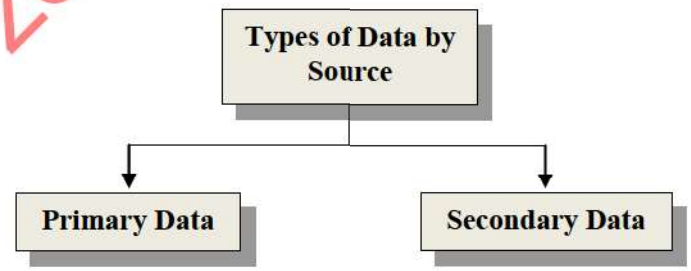
"Originally collected observations are collectively called data".

- Data of selected student's Names: Ajmal, Arif and Ali etc.
- Data of selected student's class numbers: 39, 56 and 47 etc.
- Data of selected student's heights: 60", 65" and 66".
- Data of selected student's ages: 19, 20 and 21 etc.
- Data of selected student's favorite color: Red, Blue and Green etc.

Names	Class No.	Heights	Age	Color
Ajmal	39	60	19	Red
Arif	56	65	20	Blue
Ali	47	66	21	Green

Types of Data by Source

- Primary Data
- Secondary Data



Primary Data

"The data that have been originally collected and have not undergone any sort of statistical treatment are called primary data. In other words, a fresh data is called a primary data".

Thus the primary data are the first hand information collected for a certain purpose.

- For example, the data in the Population Census Reports are primary because these are originally collected by the Population Census Organization.

Methods of Collection of Primary Data

Following methods are used for collection of Primary data:

- **Direct Personal Observation:** In this method, the investigator collects the information directly from the source concerned. The investigator must be qualified and experienced in related field of study and should put simple questions in a simple language, which could be answered easily.

- **Indirect Personal Investigations:** Sometimes, the informants would not either disclose the facts at all or would give wrong information. For example, the businessmen do not disclose their true incomes to the income tax authorities. In such a situation, information is collected from the third party.

- **Registration:** In this method, the information is reported to the appropriate authority. For example, the births and deaths are registered with the Municipal Committee or Corporation in urban areas and the Union Council in rural areas.

- **Estimates through Local Correspondents:** There is no formal collection of data in this method. Local agents or correspondents send the required information using their own judgments. It is a timesaving method and does not cast much. It is, however, a subjective method and gives only the estimates.

- **Investigation through Enumerators:** In this method information is collected through trained enumerators. The investigators get the forms of inquiry (called schedules) filled in from the informants. They help the informants in filling in the schedules correctly. This method is considered to be very accurate and timesaving. It is used in large-scale government inquiries like Population Census.

- **Mailed Questionnaire Method:** In this method, questionnaires along with a letter of request are sent by mail to the informants. The informants fill in the questionnaires and return them to the investigator. This method is very cheap and timesaving. But sometimes, the questionnaires are returned incomplete or full of errors.

Secondary Data

"The data that have undergone any sort of treatment by statistical methods, at least once i.e. the data have been collected, classified, tabulated or presented in some form for a certain purpose, are called secondary data".

- For example, the data in the Economic Survey of Pakistan are secondary because the Federal Bureau of Statistics, the State Bank of Pakistan, the Central Board of Revenue, etc. originally collect these.

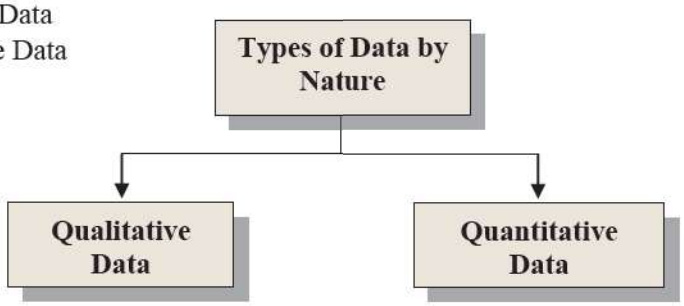
Methods of Collection of Secondary Data

Secondary data can be obtained from the following sources:

- International Publications: e.g. the publication of World Banks, I.M.F, UNESCO, UNO, ILO, etc.
- Official (Government) Sources: e.g. publication of Federal Bureau of Statistics, Ministries of Agriculture, Finance, Communications and Railways, Provincial Bureaus of Statistics and Provincial Department of Agriculture, Health and Education.
- Semi-official (Semi- Government) Sources: e.g. publications of State Bank of Pakistan, Central Cotton Committee, Economic Research Institutes, District Councils, Municipal Committees, WAPDA, P.I.D.C, etc.
- Private (Non-Government) Sources: e.g. publications of Trade Associations, Chambers of Commerce and Industry, Market Committees, etc.
- Publications of Research Organizations: e.g. Punjab University Institute of Education and Research, Irrigation Research Institute, Punjab University Social Sciences Research Center, Pakistan Institute of Development Economics, etc.

Types of Data by Nature

- Qualitative Data
- Quantitative Data



Qualitative Data

"Data collected by a qualitative variable is called qualitative data"

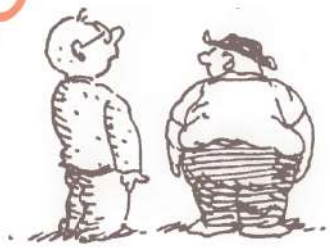
- Color
- Religion
- Gender (Female and Male)
- Education level
- Grades of students in a class etc.

Female Male

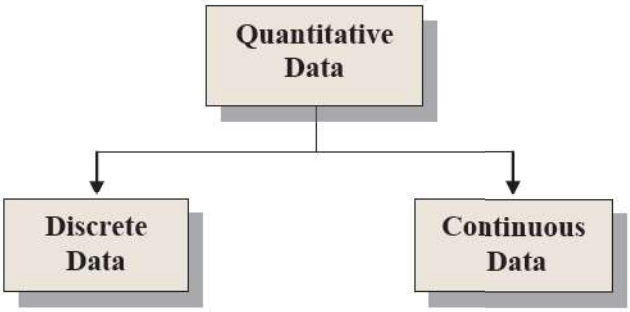
Quantitative Data

"Data collected by a quantitative variable is called quantitative data"

- Age
- Weight
- Height
- Speed
- Income
- Number of children in a family
- Number of deaths in an accident etc.



50 kg 80 kg



Historical Note

Karl Pearson

Francis Galton

When data were first analyzed statistically by Karl Pearson and Francis Galton, almost all were continuous data. In 1899, Pearson began to analyze discrete data. Pearson found that some data, such as eye color, could not be measured, so he termed such data as qualitative data.

Discrete Data

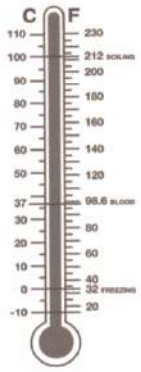
"Data collected by a Discrete variable is called discrete data"

- Family sizes
- No. of pages in a book
- No. of apples in a basket.
- No. of deaths in an accident
- No. of shares sold every day in the stock market.
- No. of housing units in different blocks of a colony
- No. of passengers carried by PIA in last ten years

Continuous Data

"Data collected by a Continuous variable is called Continuous data"

- Students heights, ages, weights
- Speed of a car
- Temperature of a place
- Income of a family
- The amount of milk given by a cow
- The life time of a TV tube
- Fortnightly petrol prices



Organization of Data

To describe situations, draw conclusions, or make inferences about events, the researcher must organize the data in some meaningful way. Thus "Organization of data means reformatting the collected data in more understandable form" The most convenient method of organizing data is **classification**, **tabulation** (frequency distributions) and constructing statistical **diagrams** and **graphs**.

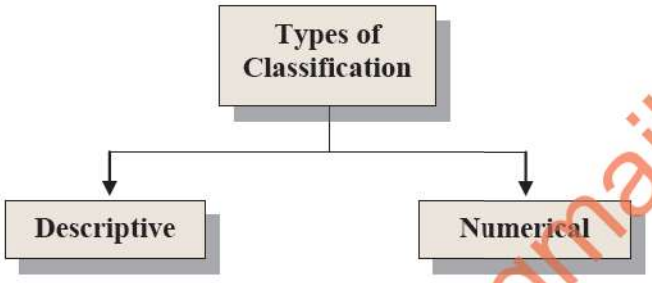
Classification

"The process of arranging data into classes or categories according to some common characteristic present in the data is called classification".

Objectives of Classification

The main objectives of classification are:

- To bring out points of similarity and dissimilarity.
- By condensing the details it saves one from mental strain.
- This enables one to make comparisons and draw inferences simply.
- It prepares the ground for the proper presentation of statistical facts.



Descriptive Classification

"When the data are classified on the basis of qualities or attributes, which are incapable of quantitative measurement, then the classification is said to be descriptive" e.g. gender, marital status, educational standard, etc. Descriptive classification is also called classification according to attributes.

Numerical Classification

"When the data are classified on the basis of quantitative measurements, then the classification is said to be Numerical" e.g. age, income, height, weights, etc.

Tabulation

"A table is a systematic arrangement of data into vertical columns and horizontal rows. Thus the process of arranging data into rows and columns is called tabulation".

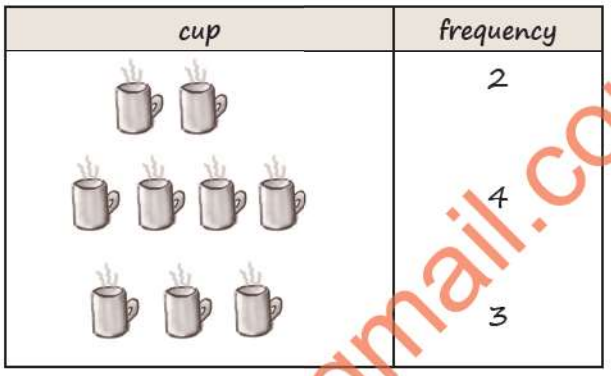
The two separate headings "Classification" and "Tabulation" should not lead the readers to assume that these are two distinct processes. Infect, they go together, classification is the first step in tabulation. Before the data are put in tabular form it has to be classified in different classes or groups having common characteristics. After this step the data are displayed under different columns and rows so that their relationship can be easily understood.

Frequency

"The number of occurrences of a particular observation in a data is called frequency".

OR

"The number of observations falling in a particular group (class) is called frequency".

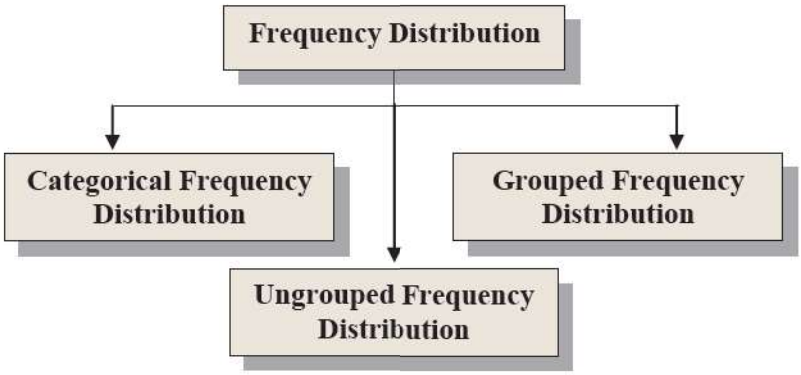


Frequency Distribution

"The organization of raw data in table form, along with frequencies is called frequency distribution".

The types of frequency distributions that will be considered here are:

- Categorical frequency distribution
- Ungrouped frequency distribution
- Grouped frequency distribution



- *A categorical frequency distribution represents data that can be placed in different categories such as gender, hair color, blood group etc. along with their frequencies.* The categorical frequency distribution is also called frequency table. A categorical frequency distribution of the students blood group is given below:

Categorical frequency Distribution (Categorical frequency Table)	
Blood Group	**No. of students *(f)***
A	5
B	8
O	4

- *An ungrouped frequency distribution simply lists the data values with the corresponding frequencies.* The ungrouped frequency distribution is also called discrete grouped data. An ungrouped frequency distribution of the students marks is given below:

Ungrouped frequency Distribution (Discrete Grouped Data)	
Marks	**No. of students *(f)***
50	4
60	6
45	3

- *The data in the form of frequency distribution is called grouped data.*
- *The purpose of a frequency distribution is to produce a meaningful pattern for the overall distribution of the data from which conclusions can be drawn.*

- *A grouped frequency distribution is obtained by constructing classes (or intervals) for the data values along with corresponding frequencies.* The grouped frequency distribution is also called continuous grouped data. A grouped frequency distribution of the heights of student is given below:

Grouped frequency Distribution (Continuous Grouped Data)	
Height	**No. of students *(f)***
3 – 4	7
4 – 5	30
5 – 6	6

Construction of frequency distribution (or frequency table) for Categorical (or Qualitative data)

Suppose that in all there are 625 students of first year in a large college. Suppose some of these students have come from Urdu medium schools and the other has come from English medium schools. If we interview the students about their schooling, we will get the observations as follows:

U, U, E, U, E, E, E, U …

(U: URDU MEDIUM) (E: ENGLISH MEDIUM)

Now the frequency table of the "medium of institution" is given as follows:

Medium of institution	No. of Students (f)
Urdu	400
English	225
Total	625

This frequency table is called **univariate frequency table** because it is constructed for one categorical variable i.e. the medium of institution.

Now suppose that along with the Medium of Institution, you are also recording the gender of the student i.e.

Student No.	Medium	Gender
1	E	F
2	E	M
3	U	F
.	.	.
.	.	.

Then the frequency table of the "medium of institution" and the "Gender of the students" is given as follows:

Medium of Institute	Gender		Total
	Male	Female	
Urdu	300	100	400
English	100	125	225
Total	400	225	625

This frequency table is called **bivariate frequency table** because it is constructed for two categorical variables i.e. the medium of institution and the gender of the students

We must decide how many categories or classes to use. These categories must be chosen so as to accommodate all the data and so that no item is placed under more than one category. The concepts of class limits, class boundaries, and class marks are of no concern when constructing frequency distribution using categorical data.

- When the data are sorted according to one criterion only is called one-way classification e.g. the student classified by the medium of institution. Tabulation in this case is called one-way tabulation.
- When the data are sorted according to two criteria is called two-way classification e.g. the student classified by the medium of institution and their gender. Tabulation in this case is called two-way tabulation.
- When the data are sorted according to three criteria is called three-way classification e.g. the student classified by the medium of institution, their gender and their residence. Tabulation in this case is called three-way tabulation.
- When the data are sorted according to many criteria is called manifold classification. Tabulation in this case is called complex tabulation.

Construction of Ungrouped Frequency Distribution (Discrete Grouped Data)

The following steps are used for constructing an ungrouped frequency distribution:

Step 1: First step is to denote the variable by X and then make a column of the X values that are in our data.

Step 2: Second step is to construct two more columns that are adjacent to the column of X. The first of these two columns is for tally marks and the second for frequency.

Step 3: Third step is to sum the frequency column and check with the total number of observations.

EXAMPLE 2.01

The following are the number of flowers on different branches of a plant:

2	4	6	1	3	3	5	7	8	6	2	9
4	7	4	2	1	3	6	4	2	5	1	4
7	9	1	2	10	1	8	9	2	3	8	2
1	2	3	4	4	4	6	6	5	5	6	1
4	5	8	5	4	3	3	2	5	0	9	1
5	9	8	10	0	10	10	--	--	--	--	--

Solution

- The variable involved is "no. of flowers" which is discrete.
- Therefore the ungrouped frequency distribution for this data is:

X (no. of flowers)	Tally	f
0	II	2
1	𝍸 III	8
2	𝍸 IIII	9
3	𝍸 II	7
4	𝍸 𝍸	10
5	𝍸 III	8
6	𝍸 II	7
7	III	3
8	IIII	4
9	𝍸	5
10	IIII	4
Total	--	67

Test Yourself

The following are the number of flowers on different branches of a plant. Construct Frequency Distribution.

12	14	16	11	13	13	15	17	18	16	12	19
14	17	14	12	11	13	16	14	12	15	11	14
17	19	11	12	20	11	18	19	12	13	18	12
11	12	13	14	14	14	16	16	15	15	16	11
14	15	18	15	14	13	13	12	15	10	19	11
15	19	18	20	10	20	20	--	--	--	--	--

Some important Points in a Grouped Frequency Distribution

Class Interval (Class)

In the following table each of the groups (110-119), (120-129) and (130,139) is called a class interval (or class).

Classes	f
110 - 119	1
120 - 129	3
130 - 139	2

Class Limits

"The smaller and larger number, which describe the class interval, are called the class limits"

- The smaller number is the **lower class limit** and the larger number is the **upper class limit**. Class limit should be well defined and there should be **no overlapping**. In other words the limits should be inclusive i.e. the values corresponding exactly to the lower limit or the upper limit be included in that class.

Classes
110 - 119
120 - 129
130 - 139
140 - 149

- Sometimes classes are taken as given in the table: In such a case, it is difficult to decide where to place an item, which is exactly 120, 130, 140, etc. because each one of them seems to belong to two classes. Such overlapping class limits should, therefore, be avoided.

Classes
110 - 120
120 - 130
130 - 140
140 - 150

- Some times a class has either no lower class limit or no upper class limit such a class is called an **open-end class**. As given in the table:

  It is clear from the above table that in the class "Below 15" there is no lower class limit and in the class "40 and over" there is no upper class limit.

Classes
Below (under) 15
15 - 19
20 - 24
25 - 29
30 - 34
35 and over (above)

Arithmetic mean, harmonic mean and geometric mean cannot be computed from an open-end frequency distribution, because the midpoints of the open-end classes cannot be determined. Therefore it is a bad practice to use open-end classes.

Class Boundaries

"The precise (true) numbers, which remove the discontinuity between two classes, are called class boundaries or true class limits"

- A class boundary is located halfway between the upper limit of a class and the lower limit of the next higher class.

Classes	Class boundaries
110 - 119	109.5 - 119.5
120 - 129	119.5 - 129.5
130 - 139	129.5 - 139.5

- If the classes are in the form:

Classes
110 - 120
120 - 130
130 - 140

Then in this case the class limits are the class boundaries because there is no discontinuity between two classes.

- Sometimes:

Classes	Class boundaries
Below (under) 15	Up to 14.5
15 - 19	14.5 - 19.5
20 - 24	19.5 - 24.5
25 - 29	24.5 - 29.5
30 - 34	29.5 - 34.5
35 - 39	34.5 - 39.5
40 and over (above)	39.5 and over

Class Mark or Mid Point

"The number, which divides each class into two equal parts, is called class mark"

It can be obtained by dividing either the sum of the lower and upper limits of a class or the sum of the lower and upper class boundaries of the class by 2.

Class Width (Class size)

"The difference between the lower class limits of two consecutive classes is called the class width". OR

"The difference between the upper class boundary and the lower class boundary of a particular class is called the class width".

The width (or size) of the class intervals is denoted by "h".

The class width may or may not be equal for all the classes. If the class width is equal for all the classes then it is called "common width". In practice it is desirable to have equal class widths whenever possible.

Construction of Grouped Frequency Distribution (Continuous Grouped Data)

The following steps are used for constructing a grouped frequency distribution:

Step 1: First step is to decide the number of classes. For this purpose there are no hard and fast rules but statistical experience tells us that no less than 5 and no more than 20 classes are generally used.

Rule: If $2^k \geq N$ then, we take "k" classes.

Where "N" is the total number of observations and "k" is the number of classes.

Step 2: Second step is to determine the range of variation in the data i.e.

$$R = X_m - X_o$$

where R is the range,
X_m is the largest value and
X_o is the smallest value.

Step 3: Third step is to determine the approximate width (size) of the class by dividing the range (R) of variation by the number of classes (k).

Step 4: Fourth step is to decide where to locate the lower class limit of the lowest class. The lowest class usually starts with the smallest data value or a number less than it (will be better if it is a multiple of class width).

Step 5: Fifth step is to list all the classes and class boundaries.

Step 6: Sixth step is to distribute the data into the appropriate classes by using a Tally-column.

Step 7: Seventh step is to complete the frequency column.

H.A Sturges has proposed an empirical rule for determining the number of classes into which a set of observations should be grouped. The rule is:

$k = 1 + 3.3 \log N$

Where k denotes the number of classes
N is the total number of observation.

A frequency distribution should have a minimum of 5 and maximum of 20 classes. For small data, use between 5 and 10 classes. For large data, use up to 20 classes.

EXAMPLE 2.02

The following data indicates number of people in different locality:

20	50	60	70	35	45	39
61	74	80	25	30	39	40
58	60	67	71	81	82	85
86	80	94	89	56	58	40
45	56	63	72	79	40	18

Solution

- The variable of given data is "number of people" which is discrete.
- Range = 94 – 18 = 76.
- Approximate no. of classes are: $2^k \geq N$
  $2^k \geq N \Rightarrow 2^6 > 35 \Rightarrow k = 6$ (N = total no. of observations)
- Class Width = Range/k = 76/6 = 12.67 ≈ 13
- Hence the grouped frequency distribution is:

Classes	Tally	f
18 – 30	IIII	4
31 – 43	𝍸 II	6
44 – 56	𝍸	5
57 – 69	𝍸 II	7
70 – 84	𝍸 IIII	9
83 – 97	IIII	4
Total	--	35

In calculating the class-width of a frequency distribution, use the next whole number as the class-width. Doing this ensures that you will have enough space in your frequency distribution for all the data values.

Alternate Method

- We may take approximate desired no. of classes = 8 (assumed)
- Class Width = Range/k = 76/8 = 9.5 ≈ 10
- Hence the grouped frequency distribution is:

Classes	Tally	f
15 – 24	II	2
25 – 34	II	2
35 – 44	𝍸 I	6
45 – 54	III	3
55 – 64	𝍸 III	8
65 – 74	𝍸	5
75 – 84	𝍸	5
85 – 94	IIII	4
Total	--	35

There is no need of Class boundaries because the variable is discrete in this case.

EXAMPLE 2.03

The following data relate to heights of 1st year students (heights in inches):

62	67	65	64	70	70	66	64	63	65
66	68	71	60	64	63	62	64	63	65
66	70	71	72	69	68	62	65	64	62
68	67	65	60	69	64	66	63	--	--

Solution

- The involved variable is "height" which is continuous variable.
- Range = 72 – 60 = 12
- No. of classes = 7 (assumed)
- Class Width =12/7 =1.714 ≈ 2
- Hence the grouped frequency distribution is:

Classes	Tally	f
60 – 62	II	2
62 – 64	𝍸 IIII	9
64 – 66	𝍸 𝍸	10
66 – 68	𝍸 I	6
68 – 70	𝍸	5
70 – 72	𝍸	5
72 – 74	I	1
Total	--	38

Here the class limits and class boundaries are the same; but it is difficult to decide where to place an item which is exactly 62, 64, and 66 etc. because each one of them seems to belong to two classes. Such overlapping class limits should therefore be avoided.

Alternate Method

Classes	Class boundaries	Tally	f
60 – 61	59.5 – 61.5	II	2
62 – 63	61.5 – 63.5	𝍸 III	8
64 – 65	63.5 – 65.5	𝍸 𝍸 I	11
66 – 67	65.5 – 67.5	𝍸 I	6
68 69	67.5 69.5	𝍸	5
70 – 71	69.5 – 71.5	𝍸	5
72 – 73	71.5 – 73.5	I	1
Total	--	--	38

Test Yourself

Construct Frequency Distribution:

1) The following data indicates number of people in different locality:

30	60	70	80	45	55	49
71	84	90	35	40	49	50
68	70	77	81	91	92	95
96	90	104	99	66	68	50
55	66	73	82	89	50	28

2) The following data relate to heights of 1st year students (heights in inches):

72	77	75	74	80	80	76	74	73	75
76	78	81	70	74	73	72	74	73	75
76	80	81	82	79	78	72	75	74	72
78	77	75	70	79	74	76	73	--	--

Relative Frequency Distribution

"The frequency of a class divided by the total of the frequencies is called the relative frequency of that class and a table showing the relative frequencies is called a relative frequency distribution".

$$R.F = \frac{frequency\ of\ a\ class}{total\ of\ frequencies\ of\ all\ classes}$$

Percentage Relative Frequency Distribution

"The frequency of a class divided by the total of the frequencies and multiplied by 100, is called the percentage relative frequency of that class and a table showing the percentage relative frequencies is called a percentage relative frequency distribution".

$$P.R.F = \frac{frequency\ of\ a\ class}{total\ of\ frequencies\ of\ all\ classes} \times 100$$

Classes	Class boundaries	f	R.F	P.R.F
60 – 61	59.5 – 61.5	2	2/38	(2/38)x100
62 – 63	61.5 – 63.5	8	8/38	(8/38)x100
64 – 65	63.5 – 65.5	11	11/38	(11/38)x100
66 – 67	65.5 – 67.5	6	6/38	(6/38)x100
68 – 69	67.5 – 69.5	5	5/38	(5/38)x100
70 – 71	69.5 – 71.5	5	5/38	(5/38)x100
72 – 73	71.5 – 73.5	1	5/38	(5/38)x100
Total	--	38	1/38	(1/38)x100
			1	100

Cumulative Frequency

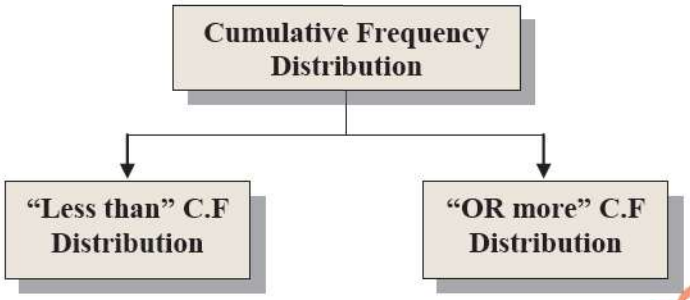
- Cumulative frequency for ungrouped frequency distribution is defined as *"total frequency that is obtained by adding the frequencies for each value to frequency for preceding values".*
- Cumulative frequency for grouped frequency distribution is defied as *"the total frequency of all classes less than the upper class boundary of a given class or the total frequency of all classes greater than the lower class boundary of given class is called cumulative frequency".*

Cumulative Frequency Distribution

"A tabular form of the variable along with the cumulative frequencies is called cumulative frequency distribution" e.g.

X	f	C.F
0	2	2
1	8	2 + 8 = 10
2	9	10 + 9 = 19
3	7	19 + 7 = 26
4	10	26 + 10 = 36
5	8	36 + 8 = 44
6	7	44 + 7 = 51
7	3	51 + 3 = 54
8	4	54 + 4 = 58
9	5	58 + 5 = 63
Total	63	--

Classes	Class boundaries	f	C.F
60 – 61	59.5 – 61.5	2	2
62 – 63	61.5 – 63.5	8	8 + 2 =10
64 – 65	63.5 – 65.5	11	11 + 10 = 21
66 – 67	65.5 – 67.5	6	6 + 21 = 27
68 – 69	67.5 – 69.5	5	5 + 27 = 32
70 – 71	69.5 – 71.5	5	5 + 32 = 37
72 – 73	71.5 – 73.5	1	1 + 37 = 38
Total	--	38	--



"Less than" Cumulative Frequency Distribution

"A "less than" cumulative frequency distribution is that, where the cumulative frequency is obtained by the total frequency of all classes less than the upper class boundary of a given class and starts with the lower class boundary of the first class indicating that there is no frequency below it".

Classes	Class boundaries	f
60 – 61	59.5 – 61.5	2
62 – 63	61.5 – 63.5	8
64 – 65	63.5 – 65.5	11
66 – 67	65.5 – 67.5	6
68 – 69	67.5 – 69.5	5
70 – 71	69.5 – 71.5	5
72 – 73	71.5 – 73.5	1
Total	--	38

Class boundaries in "less than" form	C.F
Less than 59.5	0
Less than 61.5	2
Less than 63.5	2 + 8 = 10
Less than 65.5	10 + 11 = 21
Less than 67.5	21 + 6 = 27
Less than 69.5	27 + 5 = 32
Less than 71.5	32 + 5 = 37
Less than 73.5	37 + 1 = 38
--	--

"OR more" OR "more than" Cumulative Frequency Distribution

"An "or more" cumulative frequency distribution is that, where the cumulative frequency is obtained by the total frequency of all classes more than the lower class boundary of a given class and ends with the upper class boundary of the last class indicating that there is no frequency above it."

Classes	Class boundaries	f
60 – 61	59.5 – 61.5	2
62 – 63	61.5 – 63.5	8
64 – 65	63.5 – 65.5	11
66 – 67	65.5 – 67.5	6
68 – 69	67.5 – 69.5	5
70 – 71	69.5 – 71.5	5
72 – 73	71.5 – 73.5	1
Total	--	38

Class boundaries in "or more" form	C.F
59.5 or more	38
61.5 or more	38 – 2 = 36
63.5 or more	36 – 8 = 28
65.5 or more	28 – 11 =17
67.5 or more	17 – 6 = 11
69.5 or more	11 – 5 = 6
71.5 or more	6 – 5 = 1
73.5 or more	0
--	--

Whenever we refer to a cumulative frequency distribution without any qualification, we always mean a "less than" type cumulative frequency distribution.

Relative Cumulative Frequency Distribution

"The cumulative frequency of a class divided by the total of frequencies is called the relative cumulative frequency and a table showing relative cumulative frequencies is called the relative cumulative frequency distribution".

$$R.C.F = \frac{cumulative\ frequency\ of\ a\ class}{total\ of\ \ frequencies\ of\ \ all\ classes}$$

Percentage Relative Cumulative Frequency Distribution

"The cumulative frequency of a class divided by the total of frequencies and multiplied by 100, is called the percentage relative cumulative frequency and a table showing percentage relative cumulative frequencies is called the percentage relative cumulative frequency distribution".

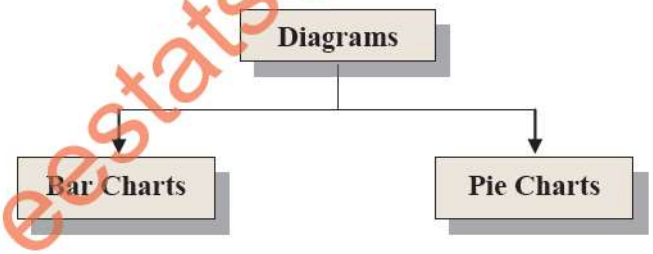
$$P.R.C.F = \frac{cumulative\ frequency\ of\ a\ class}{total\ of\ \ frequencies\ of\ \ all\ classes} \times 100$$

Classes	Class boundaries	f	C.F	R.C.F	P.R.C.F
60 – 61	59.5 – 61.5	2	2	2/38 = 0.0526	(2/38) x100 = 5.2632
62 – 63	61.5 – 63.5	8	8 + 2 =10	10/38 = 0.2632	(10/38) x100 = 26.3158
64 – 65	63.5 – 65.5	11	11 + 10 = 21	21/38 = 0.5526	(21/38) x100 = 55.2632
66 – 67	65.5 – 67.5	6	6 + 21 = 27	27/38 = 0.7105	(27/38) x100 = 71.0526
68 – 69	67.5 – 69.5	5	5 + 27 = 32	32/38 = 0.8421	(32/38) x100 = 84.2105
70 – 71	69.5 – 71.5	5	5 + 32 = 37	37/38 = 0.9737	(37/38) x100 = 97.3684
72 – 73	71.5 – 73.5	1	1 + 37 = 38	38/38 = 1	(38/38) x100 = 100
Total	--	38	--	--	--

Diagrams (Charts)

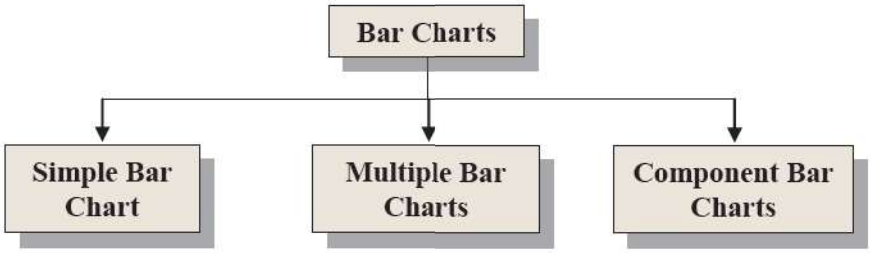
Charts or diagrams give visual representations of the **qualitative** data. Diagrams also show comparisons between two or more sets of qualitative data. Diagrams should be clear and easy to read and understand. Too much information should not be shown in the same diagram otherwise it might become confusing.

- Bar Charts
- Pie Chart or Circle Diagram



Bar Charts

- Simple bar chart
- Multiple bar charts or cluster chart
- Component bar chart or subdivided bar charts or staked bar charts



Simple Bar Chart

This chart consists of vertical or horizontal bars of equal width. The length of the bars represents the magnitude of the values of the variable i.e. the lengths of the bars vary depending on the size of data values.

EXAMPLE 2.04

The following table gives the population of five different cities. Draw a simple bar chart:

Cities	A	B	C	D	E
Population	30	45	60	63	69

Solution

Step 1: Draw the X and Y axis and place the population on Y axis.
Step 2: Draw the bars corresponding to the population.

Historical Note

In 1786 William Playfair introduced the Bar chart.

Test Yourself

The following table gives the population of six different cities. Draw a simple bar chart:

Years	A	B	C	D	E	F
Population	50	60	70	80	90	100

Multiple Bars Chart

By multiple bars chart, two or more sets of inter-related data are represented. The technique of simple bar chart is used to draw this chart but the difference is that we use different shades, colors or dots to distinguish between different phenomena. Multiple bars chart facilities comparison between more than one phenomenon.

EXAMPLE 2.05

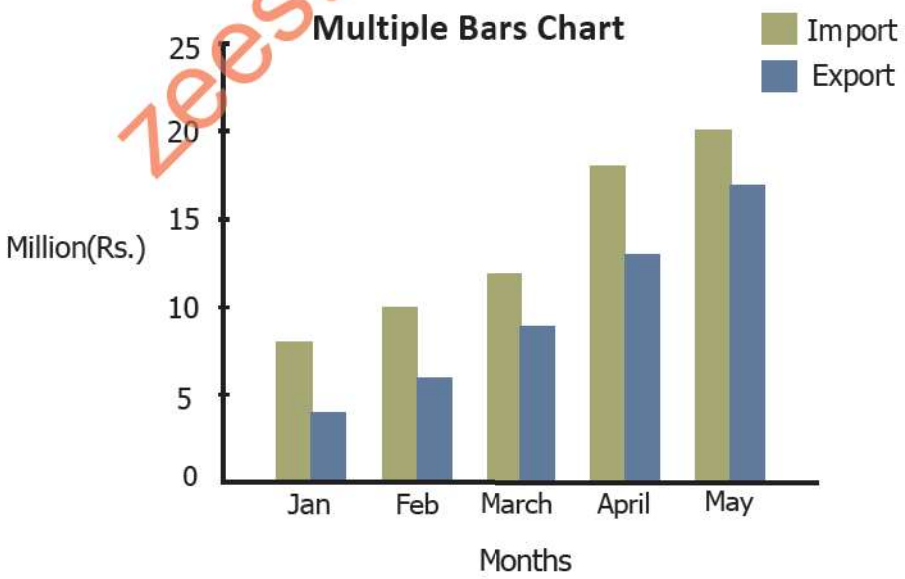
The following table gives the imports and exports of Pakistan for the five Months. Draw a multiple bar chart:

Months	Imports	Exports
January	8	4
February	10	6
March	12	9
April	18	13
May	20	17

Solution

Step 1: Draw the X and Y axis and place the amount on Y axis.

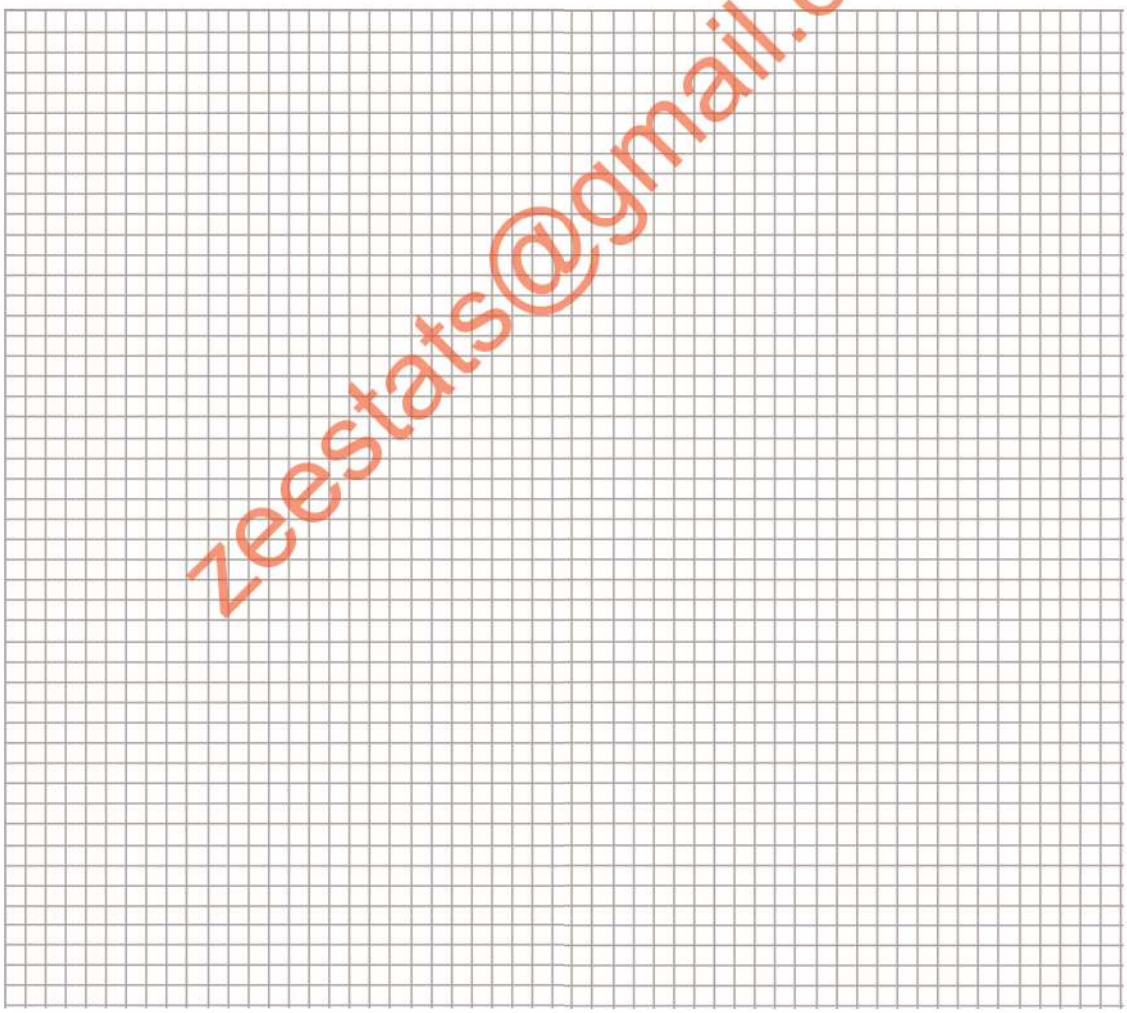
Step 2: For each month, draw two bars (both for the imports and exports) side-by-side corresponding to the amount.



Test Yourself

The following table gives the imports and exports of Pakistan for the five Months. Draw a multiple bar chart:

Months	Imports	Exports
January	9	5
February	13	9
March	15	8
April	14	16
May	23	12



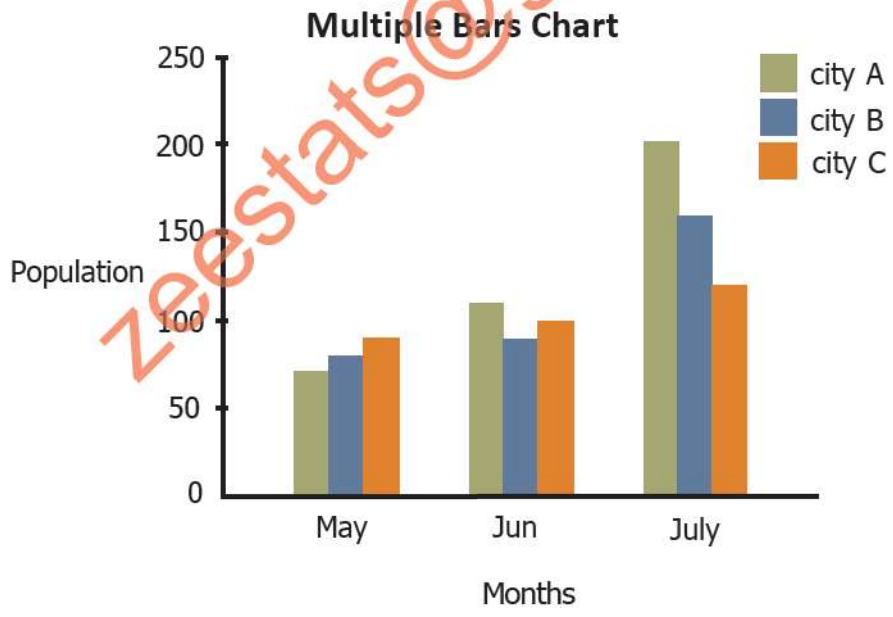
EXAMPLE 2.06

Construct a Multiple Bar Chart to show the population of the cities given in the following table:

City	Population in thousand		
	May	Jun	July
A	70	110	200
B	80	90	160
C	90	100	120

Solution

Step 1: Draw the X and Y axis and place the population on Y axis.
Step 2: For each year, draw three bars (for the three cities) side-by-side corresponding to the populations.



Test Yourself

Construct a Multiple Bar Chart to show the population of the cities given in the following table:

City	Population in thousand		
	Jan	Feb	March
A	90	120	180
B	60	80	200
C	70	120	100

Sub-divided Bar Chart

A sub-divided bar chart is an effective technique in which each bar is sub-divided into two or more parts. The component parts are shaded or colored differently to increase the overall effectiveness of the diagram.

EXAMPLE 2.07

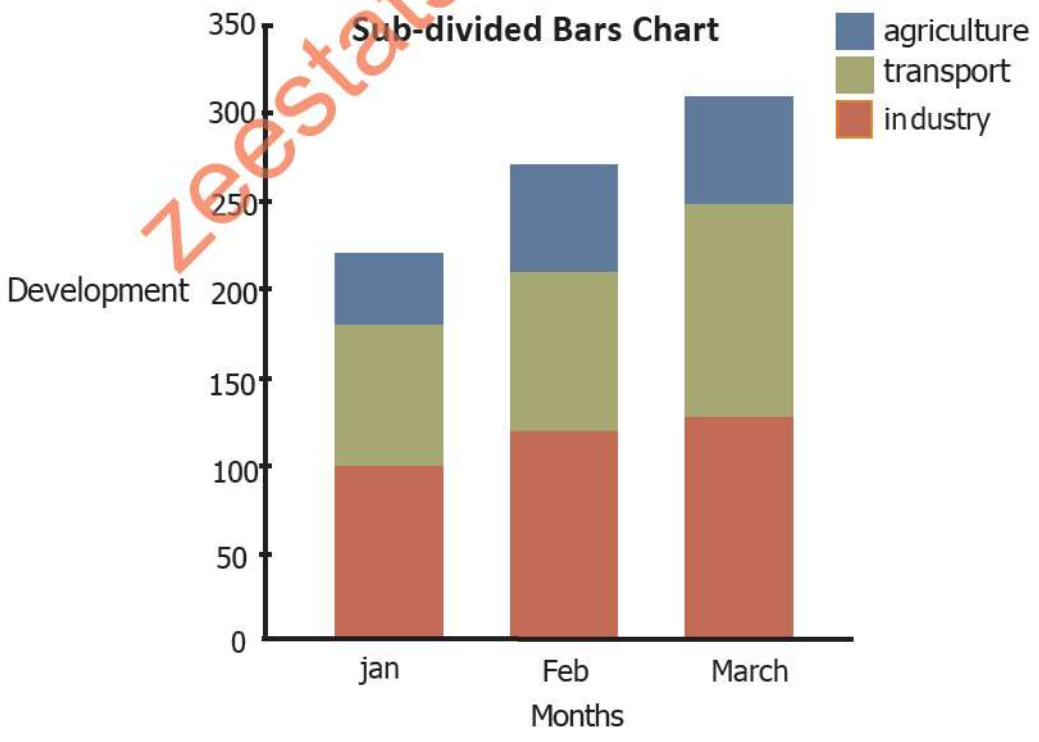
The following table represents the monthly development in the filed of industry, transport and agriculture of Pakistan. Construct a Sub-divided Bar Chart:

Months	Industry	Transport	Agriculture	Total
Jan	100	80	40	220
Feb	120	100	50	270
March	130	120	60	310

Solution

Step 1: Draw the X and Y axis and place the Development on Y axis.

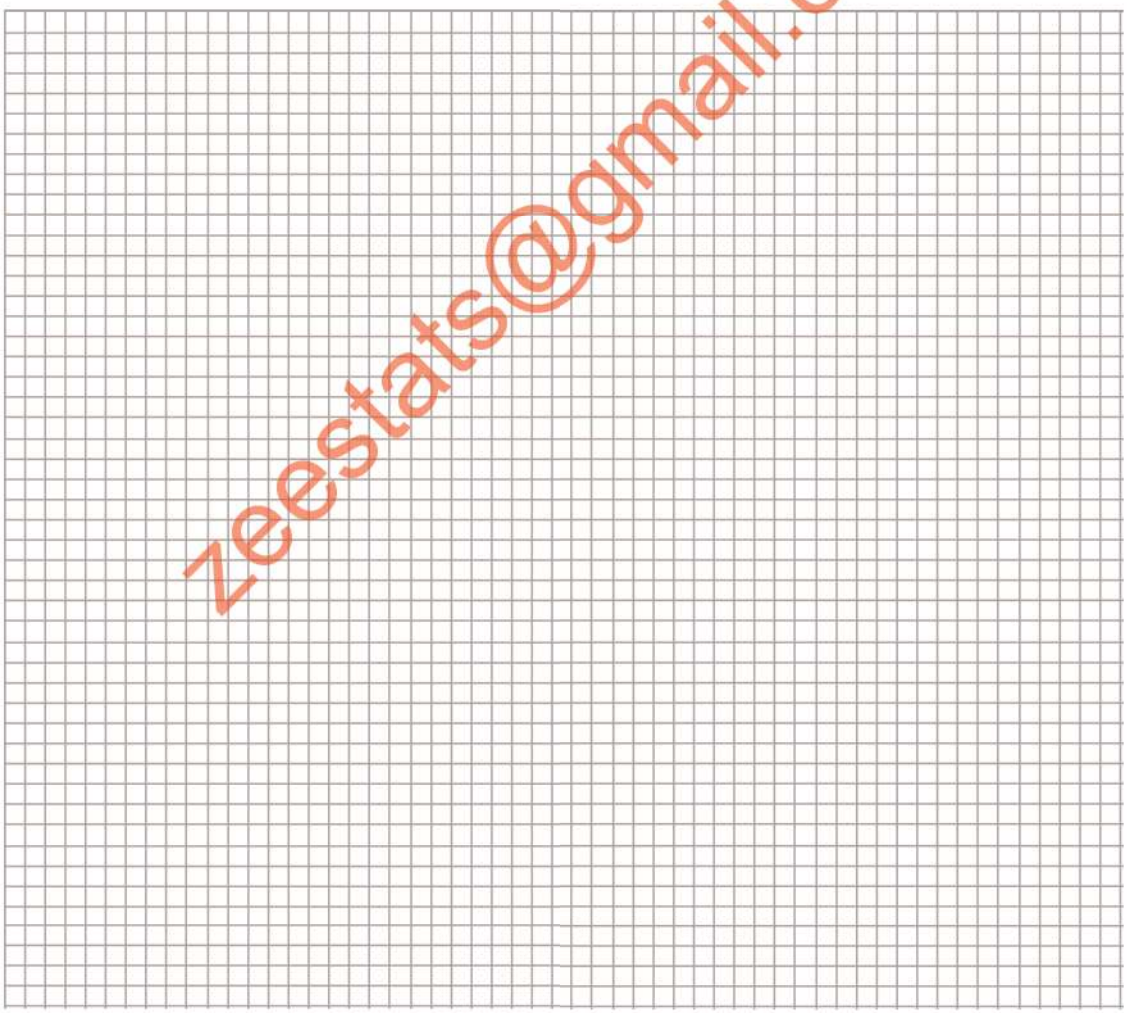
Step 2: For each year, draw three bars corresponding to the development then sub-divide each bar for the agriculture, transport and industry by their corresponding Developments.



Test Yourself

The following table represents the Monthly development in the filed of industry, transport and agriculture of Pakistan. Construct a Sub-divided Bar Chart:

Years	Industry	Transport	Agriculture	Total
May	120	90	50	260
June	140	110	70	320
July	150	100	40	290



Pie Chart or Circle Diagram

A pie-diagram, also known as sector or circle diagram, is a device consisting of a circle divided into sectors or pie-shaped pieces whose areas are proportional to the various parts into which the whole quantity is divided. The sectors are shaded or colored differently.

Historical Note

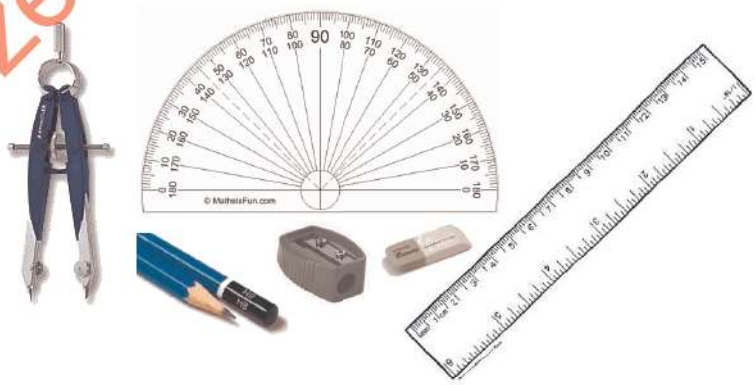
In 1801 the earliest known Pie chart is generally credited to William Playfair.

The procedure of constructing a pie chart is very simple:

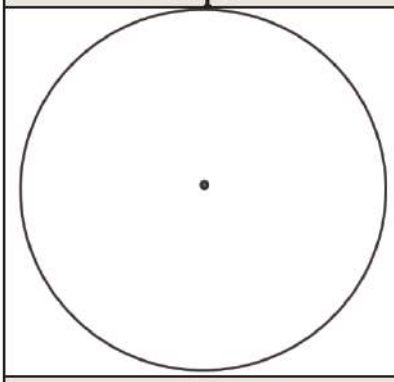
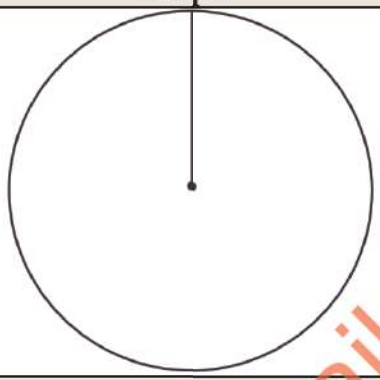
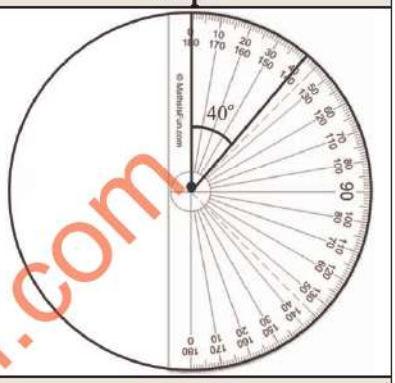
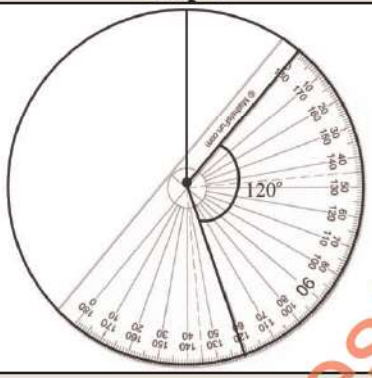
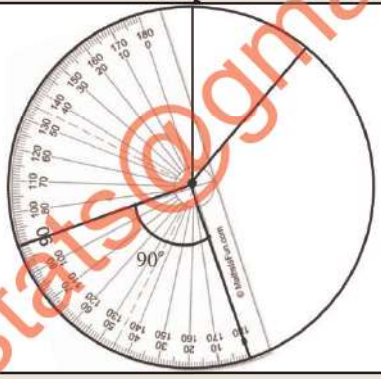
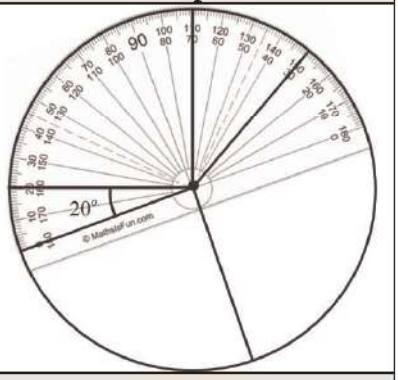
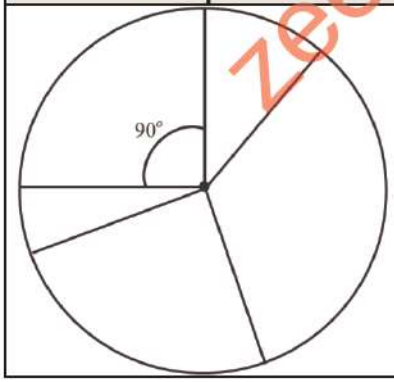
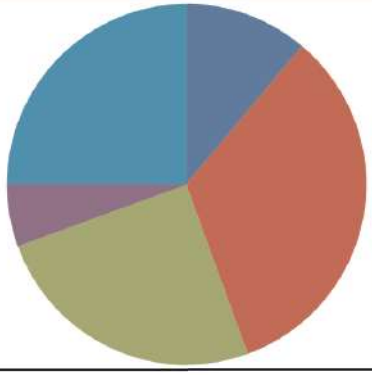
- Draw a circle of some suitable radius.
- As a circle consists of 360^{o}, the whole quantity to be displayed is equated to 360.
- Then divide the circle into different sectors by constructing angles at the center by means of a protractor and draw the corresponding radii.
- The angles are calculated by the following formula:

$$Angle = \frac{Component\ Part}{Whole\ Quantity} \times 360^{o}$$

The following **tools** are used to draw the **pie chart**.



How to draw Pie chart?

Step 1	Step 2	Step 3
		
Step 4	**Step 5**	**Step 6**
		
Step 7	**Step 8**	
		**Wow!**

EXAMPLE 2.08

Draw a Pie-diagram for the following data:

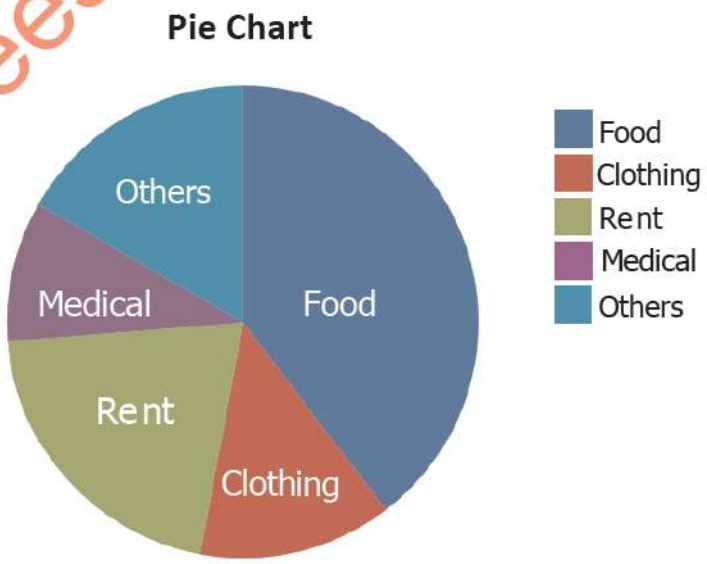
Items	Expenditure in Rs.
Food	190
Clothing	64
Rent	100
Medical	46
Other	80

Solution

Step 1: To construct Pie-diagram; first we find Angles.

Items	Expenditure in Rs.	$Angle = \frac{Component\ Part}{Whole\ Quantity} \times 360^0$
Food	190	142.5
Clothing	64	48
Rent	100	75
Medical	46	34.5
Other	80	60
Total	480	360

Step 2: Next, using a protractor and a compass, draw the graph using the appropriate degree found in step 1, and label each section with the name.



Test Yourself

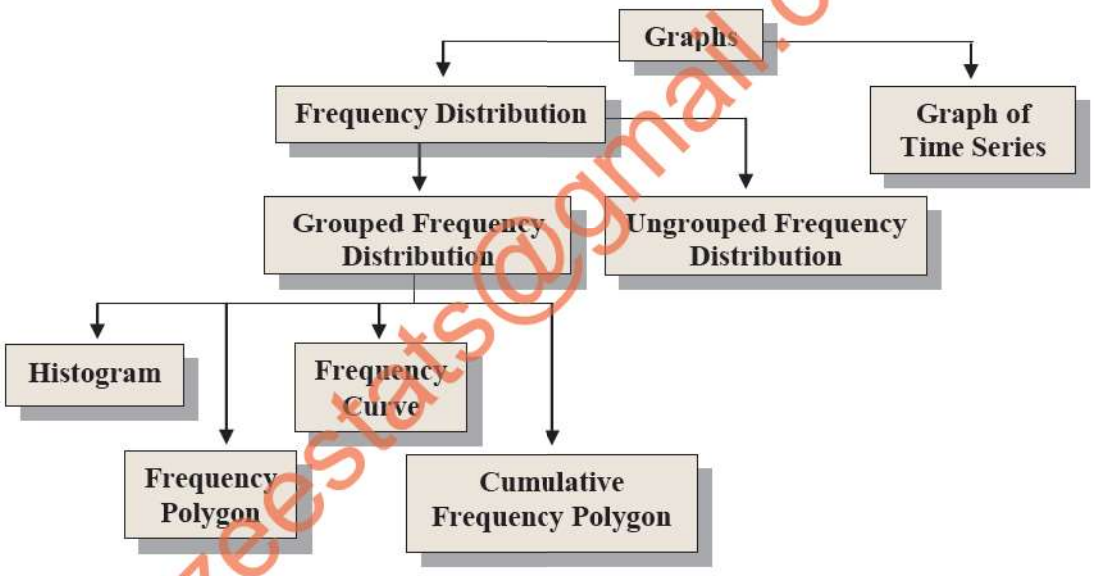
Draw a Pie-diagram for the following data:

Items	Expenditure in Rs.
Food	160
Clothing	80
Rent	120
Medical	50
Other	90

Graphs

Graphs give visual representations of the **quantitative** data. A graph consists of curves or straight lines. Graphs provide a very good method of showing fluctuations and trends in statistical data. Graphs can also be used to make predictions and forecasts.

- Histogram
- Frequency Polygon
- Frequency Curve
- Cumulative Frequency Polygon (Ogive)
- Graph of Ungrouped Frequency Distribution
- Graph of **Time Series**



Histogram

A histogram consists of a set of rectangles having bases on a horizontal axis i.e. X-axis (note that these bases are marked off by class boundaries not class limits) with centers at the class marks and areas proportional to the class frequencies.

- If the widths of the classes are equal then the heights of the rectangles are also proportional to the class frequencies and are taken numerically equal to class frequencies.
- If the widths of the classes are not equal then the heights of the rectangles have to be adjusted.

Historical Note

In 1891, Pearson introduced "Histogram"

First Method (Equal Class Width)

- Draw X-axis and Y-axis.
- Take class boundaries on X-axis and frequencies on Y-axis.
- Construct joint rectangles. The resulting figure is the required histogram.

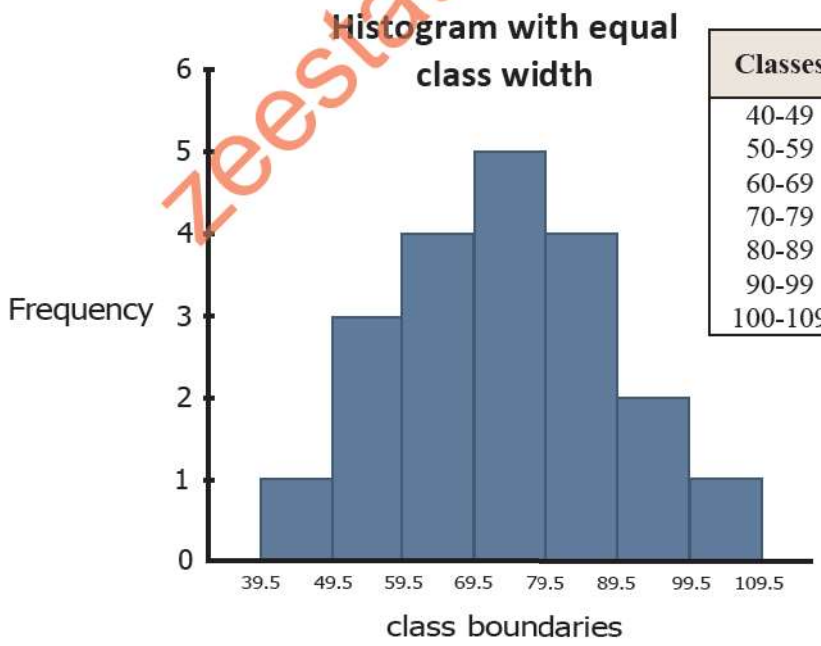
EXAMPLE 2.09

Construct Histogram from the following frequency distribution:

Classes	40-49	50-59	60-69	70-79	80-89	90-99	100-109
Frequency	1	3	4	5	4	2	1

Solution To draw a Histogram we proceed with the following steps:

Step 1: Find class-boundaries.
Step 2: Mark class-boundaries along the x-axis and the frequencies along y-axis.
Step 3: Construct rectangles having width proportional to class widths and heights proportional to class frequencies.
Step 4: The resulting graph will be the Histogram as given below.



Classes	f	Class-boundaries
40-49	1	39.5-49.5
50-59	3	49.5-59.5
60-69	4	59.5-69.5
70-79	5	69.5-79.5
80-89	4	79.5-89.5
90-99	2	89.5-99.5
100-109	1	99.5-109.5

Test Yourself

Construct Histogram from the following frequency distribution:

Classes	30-39	40-49	50-59	60-69	70-79	80-89	90-99
Frequency	2	4	5	7	4	3	2

Second Method (Unequal Class Width)

- Draw X-axis and Y-axis.
- Take class boundaries on X-axis and adjusted frequencies on Y-axis. (Frequencies are adjusted by dividing them by their respective class width)
- Construct joint rectangles. The resulting figure is the required histogram.

EXAMPLE 2.10

Construct Histogram from the following frequency distribution:

classes	40-49	50-53	54-64	65-79	80-89	90-99	100-109
f	10	12	44	75	40	20	10

Solution To draw a Histogram we proceed with the following steps:

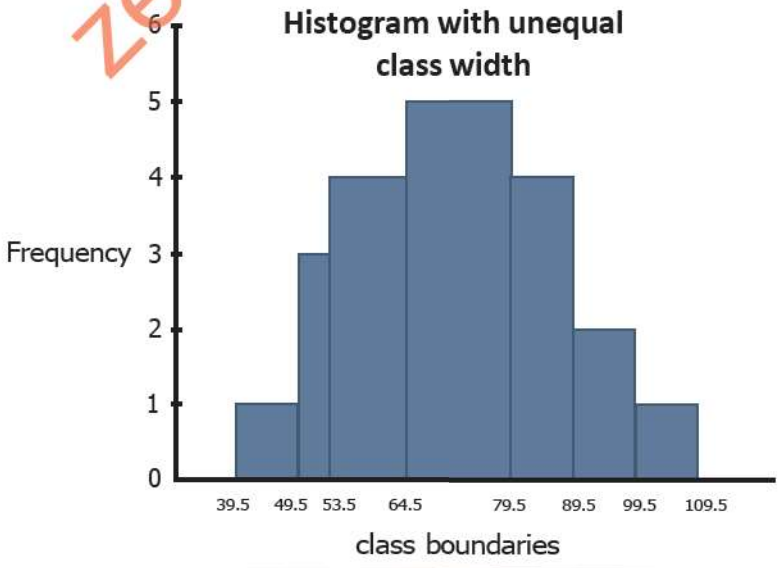
Step 1: Find class-boundaries and adjusted frequencies.

Step 2: Mark class-boundaries along the x-axis and the adjusted frequencies along y-axis.

Step 3: Construct rectangles having width proportional to class-width and heights proportional to class adjusted frequencies.

Step 4: The resulting graph will be the Histogram as given below.

Class-boundaries	Class width (h)	$Adj: frequency = \frac{f}{h}$
39.5-49.5	10	1
49.5-53.5	4	3
53.5-64.5	11	4
64.5-79.5	15	5
79.5-89.5	10	4
89.5-99.5	10	2
99.5-109.5	10	1



Test Yourself

Construct Histogram from the following frequency distribution:

classes	20-29	30-33	34-44	45-59	60-69	70-79	80-89
f	12	15	48	80	30	25	15

Frequency Polygon

A frequency polygon is a many sided closed figure. It is constructed by plotting the class frequencies against their corresponding class marks (mid-points) and then joining the resulting points by means of straight lines. It can also be obtained by joining the mid-points of the tops of rectangles in the histograms.

Method

- Draw X-axis and Y-axis.
- Take class marks on X-axis and frequencies on Y-axis.
- Join the points by means of straight lines. The resulting figure is the required frequency polygon.

In this method the ends of the graph do not meet the X-axis and we know that a polygon is a many-sided closed figure. We may therefore add extra classes at both ends of frequency distribution with zero frequencies. By doing so the polygon forms a closed figure.

EXAMPLE 2.11

Construct Frequency Polygon from the following frequency distribution:

Classes	10-19	20-29	30-39	40-49	50-59
Frequency	5	15	40	20	10

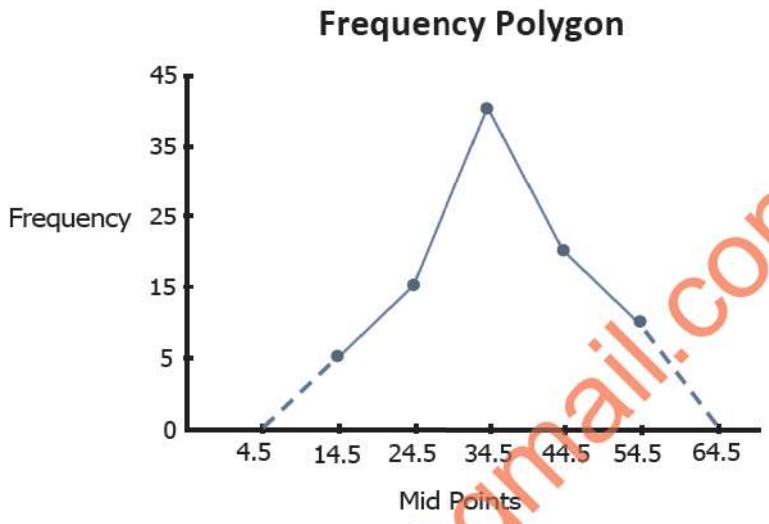
Solution

To draw a Frequency Polygon we proceed with the following steps:

Step 1: Find class-marks (mid-points).

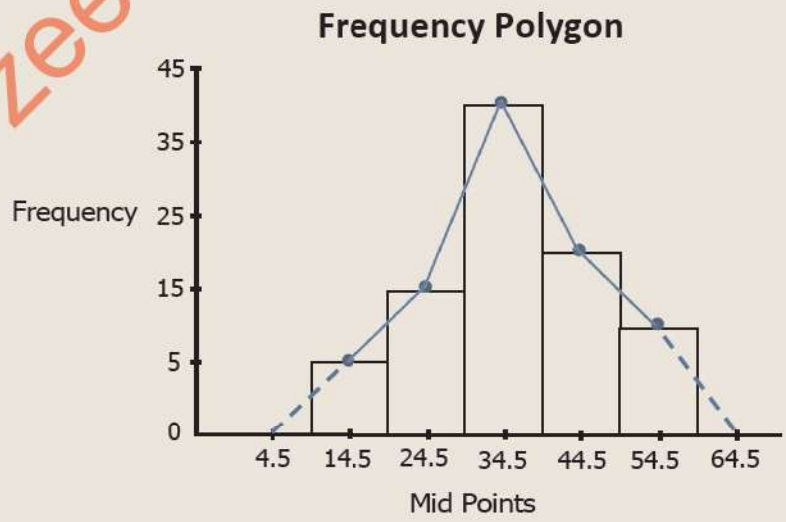
Classes	Frequency	Mid-points
10-19	5	14.5
20-29	15	24.5
30-39	40	34.5
40-49	20	44.5
50-59	10	54.5

Step 2: Mark mid-points along the x-axis and the frequencies along y-axis.
Step 3: Place a dot against each mid-point with respect to its class frequency.
Step 4: Join the dots by straight lines to get Frequency Polygon as given below.



We can also draw a frequency polygon by the following method:

- *Draw a histogram*
- *The mid-points at the top of each rectangle are joined by straight lines. The figure is the required frequency polygon.*



Test Yourself

Construct Frequency Polygon from the following frequency distribution:

Classes	20-29	30-39	40-49	50-59	60-69
Frequency	7	13	42	23	6

Frequency Curve

When the frequency polygon is smoothed out as a curve then it becomes frequency curve. OR when the mid-points are potted against the frequencies then a smooth curve passes through these points is called a frequency curve.

Method

- Draw X-axis and Y-axis.
- Take class marks on X-axis and frequencies on Y-axis.
- Plot the frequencies against the class marks.
- The plotted points are then joined by a smooth curve, which gives frequency curves.

EXAMPLE 2.12

Construct Frequency Curve from the following frequency distribution:

Classes	10-19	20-29	30-39	40-49	50-59
Frequency	5	15	40	20	10

Solution

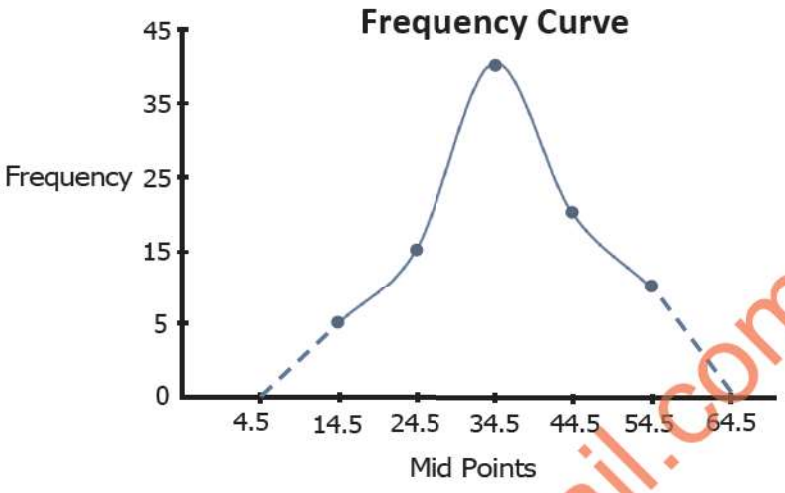
To draw a Frequency Curve we proceed with the following steps:

Step 1: Find class-marks (mid-points).

Classes	Frequency	Mid-points
10-19	5	14.5
20-29	15	24.5
30-39	40	34.5
40-49	20	44.5
50-59	10	54.5

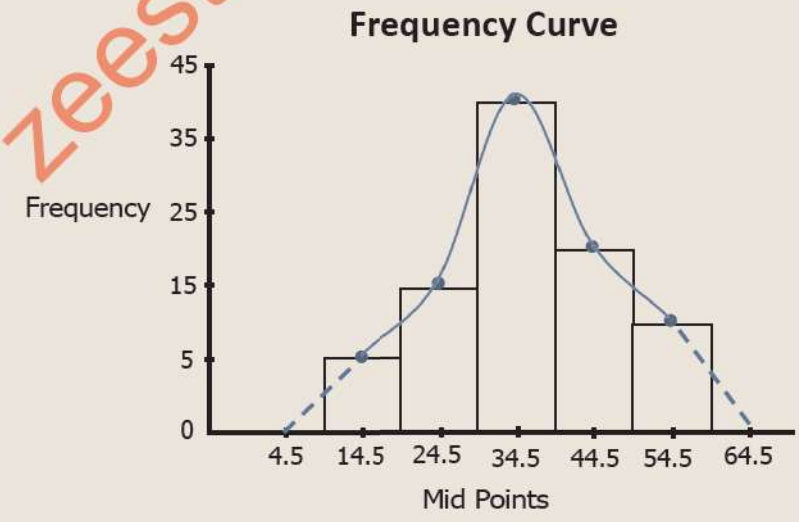
The smoothed curve should pass above the highest points of the polygon

Step 2: Mark mid-points along the x-axis and the frequencies along y-axis.
Step 3: Place a dot against each mid-point with respect to its class frequency.
Step 4: Join the dots by smooth line to get Frequency Curve as given below.



We can also draw a frequency curve by the following method:

- Draw X-axis and Y-axis.
- Draw a histogram.
- Draw a smooth curve through the top of the rectangles. The resulting figure is the required frequency curve.



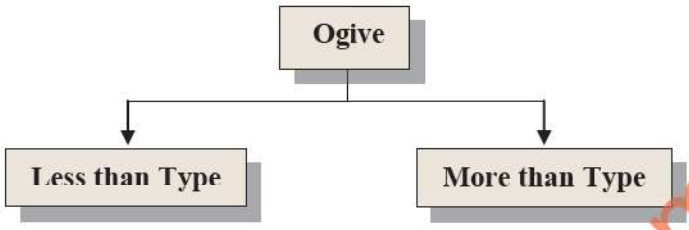
Test Yourself

Construct Frequency Curve from the following frequency distribution:

Classes	20-29	30-39	40-49	50-59	60-69
Frequency	6	16	45	18	8

Cumulative Frequency Polygon (Ogive)

When a curve is based on cumulative frequencies then it is called an ogive.



Less than Type

Method

- First calculate the cumulative frequencies.
- Take upper class boundaries on X-axis and the cumulative frequencies on Y-axis.
- Plot the cumulative frequency against the upper class boundaries.
- Join the potted points by straight lines. The resulting figure is the required less than cumulative frequency polygon or less than ogive.

EXAMPLE 2.13

Construct Less Than Cumulative Frequency Polygon (Ogive) from the following frequency distribution:

Classes	10-19	20-29	30-39	40-49	50-59
Frequency	5	25	45	15	10

Solution To draw a Cumulative Frequency Polygon (Ogive) we proceed with the following steps:

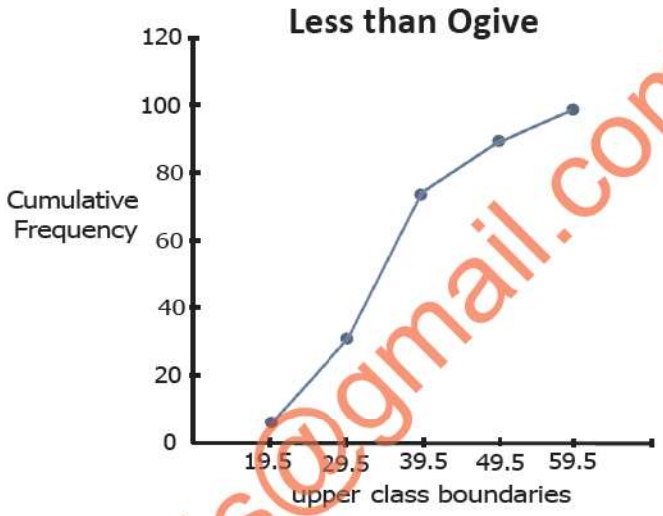
Step 1: Find class-boundaries and cumulative frequencies.

Classes	Frequency	Cumulative Frequency	Class boundaries
10-19	5	5	9.5-19.5
20-29	25	30	19.5-29.5
30-39	45	75	29.5-39.5
40-49	15	90	39.5-49.5
50-59	10	100	49.5-59.5

Step 2: Mark upper class-boundaries along the x-axis and cumulative frequencies along y-axis.

Step 3: Place a dot against each upper class-boundary with respect to its class cumulative frequency.

Step 4: Join the dots by straight line to get Cumulative Frequency Polygon (Ogive) as given below.



More than Type

Method

- First calculate the cumulative frequencies.
- Take lower class boundaries on X-axis and the cumulative frequencies on Y-axis.
- Plot the cumulative frequency against the lower class boundaries.
- Join the potted points by straight lines. The resulting figure is the required more than cumulative frequency polygon or more than ogive.

If we join the points in cumulative frequency polygon by smoothed line then we get a smoothed ogive.

EXAMPLE 2.14

Construct More Than Cumulative Frequency Polygon (Ogive) from the following frequency distribution:

Classes	10-19	20-29	30-39	40-49	50-59
Frequency	5	25	45	15	10

Solution To draw a Cumulative Frequency Polygon (Ogive) we proceed with the following steps:

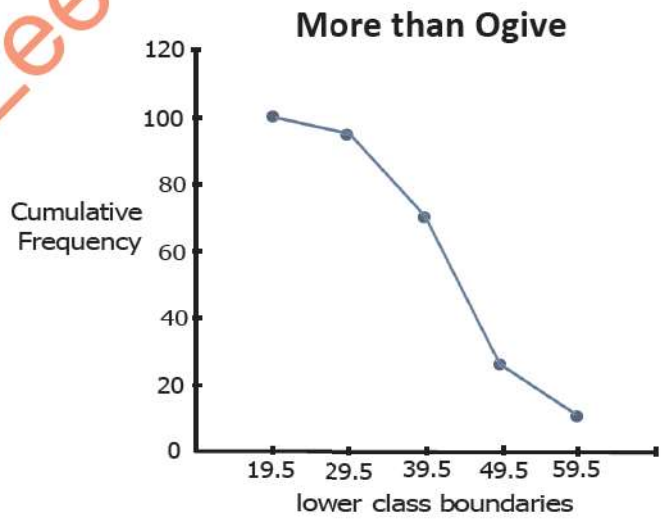
Step 1: Find class-boundaries and cumulative frequencies.

Classes	Frequency	Cumulative Frequency	Class boundaries
10-19	5	100	9.5-19.5
20-29	25	95	19.5-29.5
30-39	45	70	29.5-39.5
40-49	15	25	39.5-49.5
50-59	1	10	49.5-59.5

Step 2: Mark lower class-boundaries along the x-axis and cumulative frequencies along y-axis.

Step 3: Place a dot against each lower class-boundary with respect to its class cumulative frequency.

Step 4: Join the dots by straight line to get Cumulative Frequency Polygon (Ogive) as given below.



Test Yourself

Construct Both More Than and Less Than Cumulative Frequency Polygon (Ogive) from the following frequency distribution:

Classes	30-39	40-49	50-59	60-69	70-79
Frequency	4	13	43	26	12

Graph of Ungrouped Frequency Distribution

Vertical lines graph is visual representation of an **ungrouped frequency distribution**. It consists of a set of vertical lines that are perpendicular to the X- axis and intersect the X-axis at the values of the **discrete variable** and the height of each line is proportional to its frequency.

First Method

- Draw X-axis and Y-axis.
- Take the values of discrete variable on X-axis and frequencies on Y-axis.
- Draw **vertical lines** for each value of the variable such that the height of each line is proportional to its frequency.

EXAMPLE 2.15

Construct vertical lines graph for the following data:

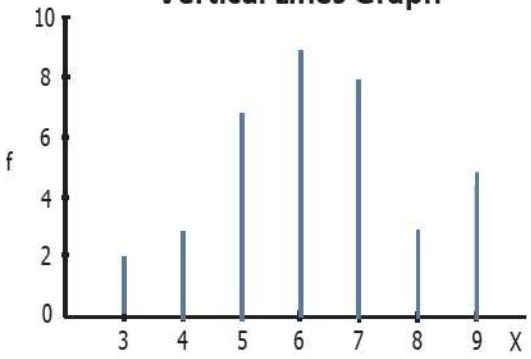
X	3	4	5	6	7	8	9
f	2	3	7	9	8	3	5

For discrete variable, if we make Histogram we first find class boundaries. These class boundaries are called factitious class boundaries because the discrete variable cannot assume such values.

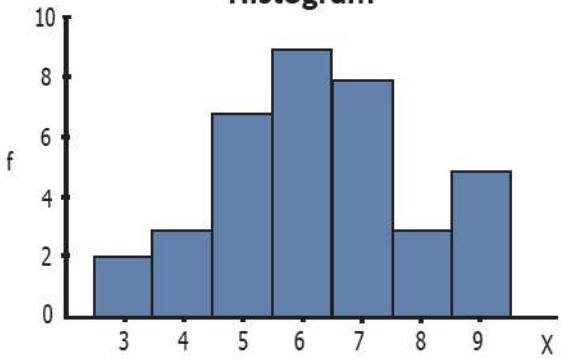
Solution

Step 1: Draw X-axis and Y-axis.
Step 2: Take the variable "X" on X-axis and frequencies along Y-axis.
Step 3: Draw vertical lines for each value of "X" with height equal to its frequency.

Vertical Lines Graph



Histogram



Test Yourself

Construct vertical lines graph and the histogram for the following data:

X	30	40	50	60	70	80	90
Frequency	4	6	7	13	12	5	1

Graph of Time Series

A curve showing changes in the value of one or more items from one period of time to the next is known as the graph of time series. Thus a Graph of time series displays the variations in time series dealing with prices, production, imports, population etc.

Method

- Draw X-axis and Y-axis.
- Take time (years, months, weeks, etc.) along X-axis and the corresponding values along Y-axis.
- Plot the various points. Join the plotted points by **straight lines**. The resulting figure is the required graph of time series.

Do not try to fit a smooth curve through the data points

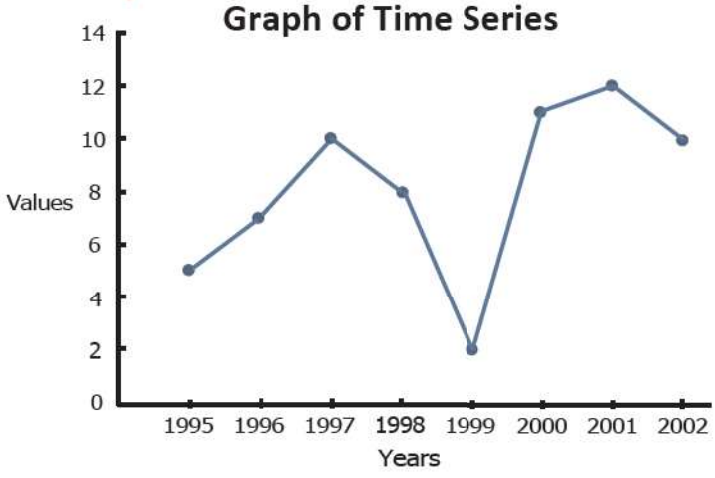
EXAMPLE 2.16

Construct Graph from the following Time Series:

Year	1995	1996	1997	1998	1999	2000	2001	2002
Values	5	7	10	8	2	11	12	10

Solution

Step 1: Draw X-axis and Y-axis.
Step 2: Take time along X-axis and the corresponding values along Y-axis.
Step 3: Plot the various points. Join the plotted points by straight lines. The resulting figure is the required graph of time series.



Historical Note

In 1786 William Playfair invented the line graph.

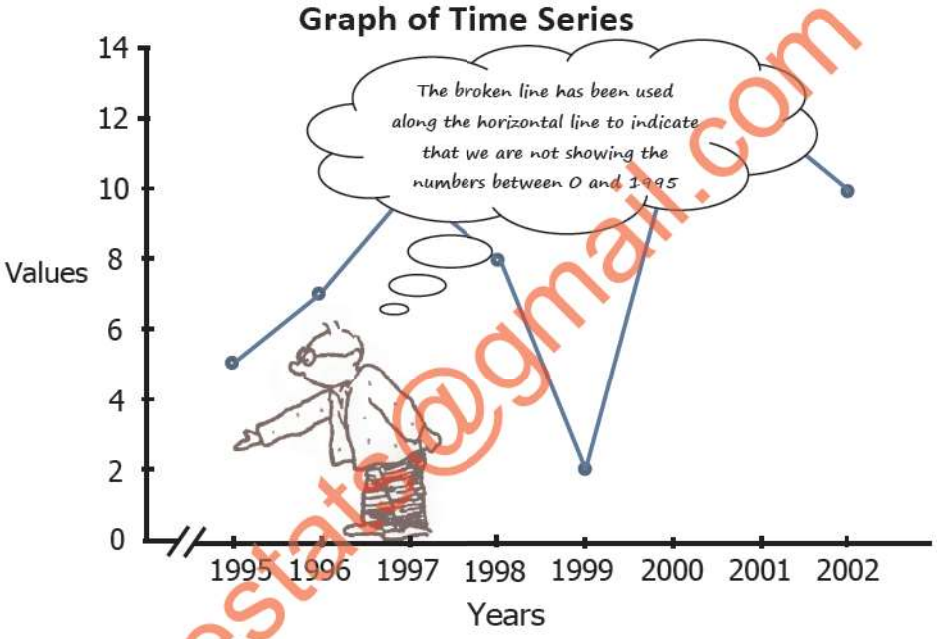
Test Yourself

Construct Graph from the following Time Series:

Year	2000	2001	2002	2003	2004	2005	2006	2007
Values	8	9	15	12	10	9	11	7

False Base Line or the broken line

In all the above graphs and diagram, if the horizontal scale is started from zero it would not only be difficult to accommodate the whole data on the graph paper but the graph would go at the right of the paper. In order to avoid this, **false base line** is used. In false base line, instead of showing the entire horizontal scale starting from zero to the highest value involved, only that portion of the scale is shown which serves the purpose. Thus the portion of the scale, starting from zero to the minimum value is omitted.



The Difference between Bar Charts and Histograms

- *Here is the main difference between bar charts and histograms. With bar charts, each column represents a group defined by a categorical variable; and with histograms, each column represents a group defined by a quantitative variable.*
- *It is always appropriate to talk about the skewness of a histogram; that is, the tendency of the observations to fall more on the low end or the high end of the X axis.*
- *With bar charts, however, the X axis does not have a low end or a high end; because the labels on the X axis are categorical – not quantitative. As a result, it is less appropriate to comment on the skewness of a bar chart.*

Following differences may be noted between diagrams and graphs.

- In the construction of a graph, graph paper is used. A graph helps to study the mathematical relation between two variables such as price and demand; income and consumption, time and population etc. On the other hand, diagrams are generally constructed on a plain paper. A diagram is used for sake of comparison but not for studying the relation between two variables.

- Graphs are more precise and accurate than diagrams. They are more helpful to a researcher for studying the relationship between two variables and for further statistical analysis and interpretation. Diagrams furnish only approximate information on the problem under study. These are not much use to a researcher for further analysis.

- Graphs are used to present time series data and frequency distributions. Diagrams are useful in presenting qualitative data. Presentation of data through graphs is easier than through diagrams.